DOCUMENT RESUME

ED 174 685 TM 009 657

AUTHOR Groome, Mary Lynn; Groome, William R.
TITLE Item Bias Identification: A Comparison of Two Approaches.
PUB DATE Apr 79
NOTE 21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 9-11, 1979); Tables marginally legible

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Complexity Level; Identification; *Item Analysis; Mathematical Models; Simulation; *Test Bias; *Test Items
IDENTIFIERS *Angoff Method; Birnbaum 3 Parameter Logistic Model

ABSTRACT

Angoff's method for identifying possible biased test items was applied to four computer-generated hypothetical tests, two of which contained no biased items and two of which contained a few biased items. The tests were generated to match specifications of a latent trait model. Angoff's method compared item difficulty estimates for two different groups, using a normal transformation of percent passing in each of the two groups for the difficulty statistics. The criteria for bias used in this study correctly identified 10 out of 21 supposedly biased items. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED174685

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ITEM BIAS IDENTIFICATION: A COMPARISON OF TWO APPROACHES

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Mary Lynn Groome
William Groome

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Mary Lynn Groome, M.A.
School of Education
California State University, Sacramento[1]

and

William R. Groome, M.A.
Department of Psychology
California State University, Sacramento[2]

A paper presented at the Annual Meeting of the National Council for Measurement in Education, San Francisco, California, April 1979.

[1]Now Research Assistant
State of California, California Energy Commission
Conservation Division
Utility Systems Office

[2]Now Research Analyst
State of California, Department of Consumer Affairs
Administrative Division
Central Testing Unit

TM009 657

ITEM BIAS IDENTIFICATION: A COMPARISON OF TWO APPROACHES

Introduction and Background

Ever since group testing has been used on a mass scale, the question of bias has been a major issue. To properly address this issue, test developers need models to determine which items within a test are biased in order to construct culturally fair tests. Several such models have been offered by various researchers. Yet, the effectiveness of these models has not been adequately tested using a common data set. Groome (1977) and Groome (1977) examined two such models; the Transformed Item Difficulties approach (TID) and the three parameter Latent Trait Model approach (LTM).

The TID approach, as first developed by Angoff (1972), defines item bias as an item by group interaction. This model consists of obtaining the p-value item difficulties for two different groups on a set of items. Each p-value is transformed into Z-scores for each group and then plotted on a bivariate graph. Ideally, the plot would be a perfect 45° line indicating the absence of item bias. When this ideal condition does not exist, one of two possible criterion is used to identify the truely biased items. The first criterion, as developed by Strassberg-Rosenberg and Donlon (1975), says that any item varying from the 45° line by greater than 1.5 times the standard deviation of the residuals is identified as a biased item. The second criterion identifies a biased item as any item which lies outside of

the 95% confidence interval.

Figure 1 represents a hypothetical example of a transformed item p-value regression scatterplot. It demonstrates the use of the second criterion in identifying biased items. The solid line represents the 45° regression line and the dotted lines indicate the linear confidence limits. Those items falling outside of the confidence limits are considered to be biased items. In this example only two items were classified as being biased.

The three parameter LTM approach defines item bias as unequal probabilities of responding correctly to an item for persons of the same ability level, but belonging to differently defined subgroups. In LTM theory the probability of an examinee responding correctly to any given item is related to a function of the examinee's underlying ability level (θ), theta, and the characteristics (parameters) of the item: a_g, discrimination; b_g, difficulty; and c_g, the psuedo guessing parameter. This model is the cumulative logistic latent trait model developed by Birnbaum (1968) (see Lord & Novick, 1974).

Essentially, the model describes items on three different continuums (item parameters) and persons on a single continuum of ability level. Thus, each item can be represented as a monotonically increasing curve (item characteristic curve or icc) with the probability of a correct response on the ordinate and the ability level, theta (θ), on the abscissa.

Figure 2 represents a hypothetical icc. An examinee with a high true ability, θ_j, has a high probability of responding correctly to item g. On the other hand, an examinee of low ability level, θ_k, has a low probability of responding correctly.

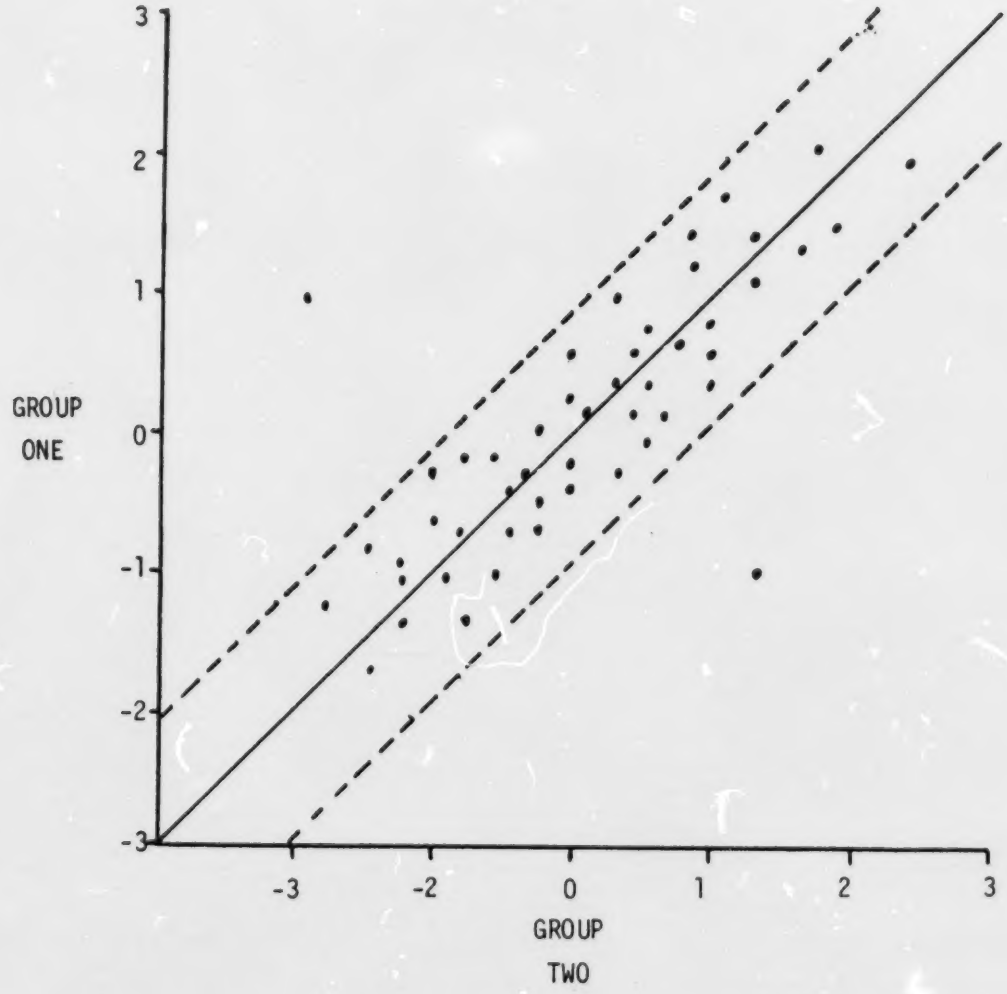


Figure 1. A hypothetical transformed p-values regression scatterplot.

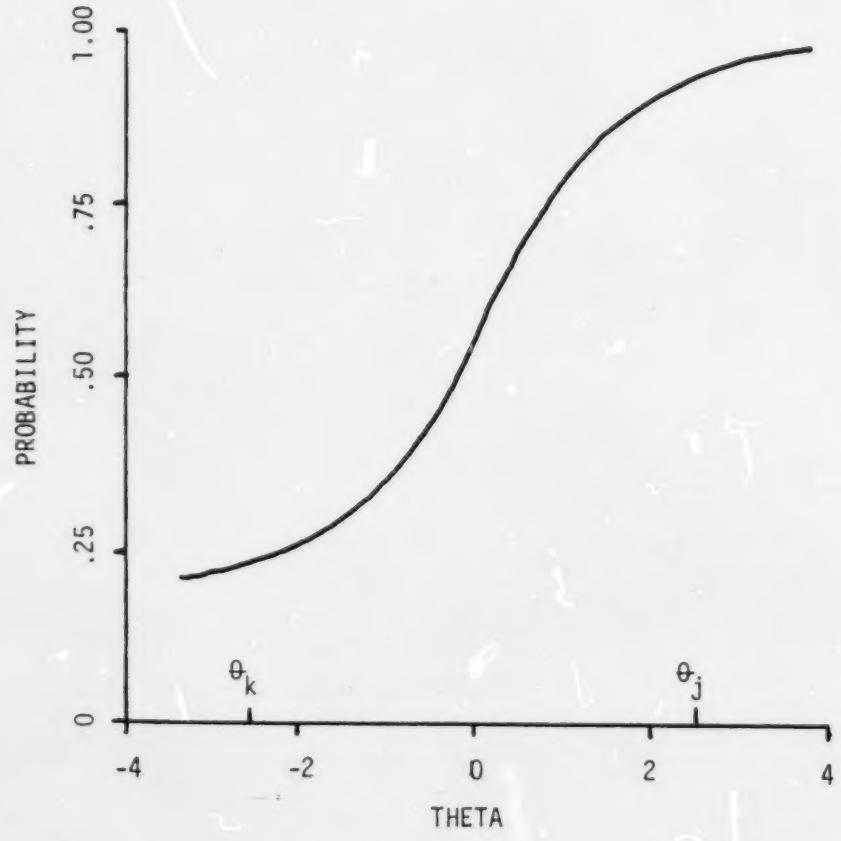


Figure 2. A hypothetical item characteristic curve.

For item bias identification, Rudner (1977), has proposed obtaining estimates of the parameters for two differently defined groups and computing the area between the two equated icc's for each group. The area represents the estimated difference in the probability of a correct response between the two groups on an item across the continuum of ability. Thus, bias is on a scale where a decision can be made as to which items are equal to or greater than a critical acceptance level. Rudner (1977) has defined items with icc area differences equal to or greater than .70 as biased items. Figure 3 represents a hypothetical example of such a biased item.

The purpose of this study was to investigate how satisfactorily the T1D approach detected known biased items: that is, items defined as biased in the computer generated tests. To accomplish our purpose, responses for two groups of examinees were computer generated using a three parameter Latent Trait Model program. Values for the parameters were choosen to create desired distributions for discrimination, difficulty, guessing, bias in difficulty and ability for each group. Both normal and uniform distributions of the item parameters were specified. Thus, four tests were developed with item responses generated randomly to simulate a 60 item test for two groups of examinees - 1000 in each group. Equal and normal distributions of ability were specified for each group on all tests. The four tests were:

TEST 1. normal distribution of item parameters with no biased items;

TEST 2. uniform distribution of item parameters with no biased items;

TEST 3. normal distribution of item parameters with several biased items; and

TEST 4. uniform distribution of item parameters with several biased items.

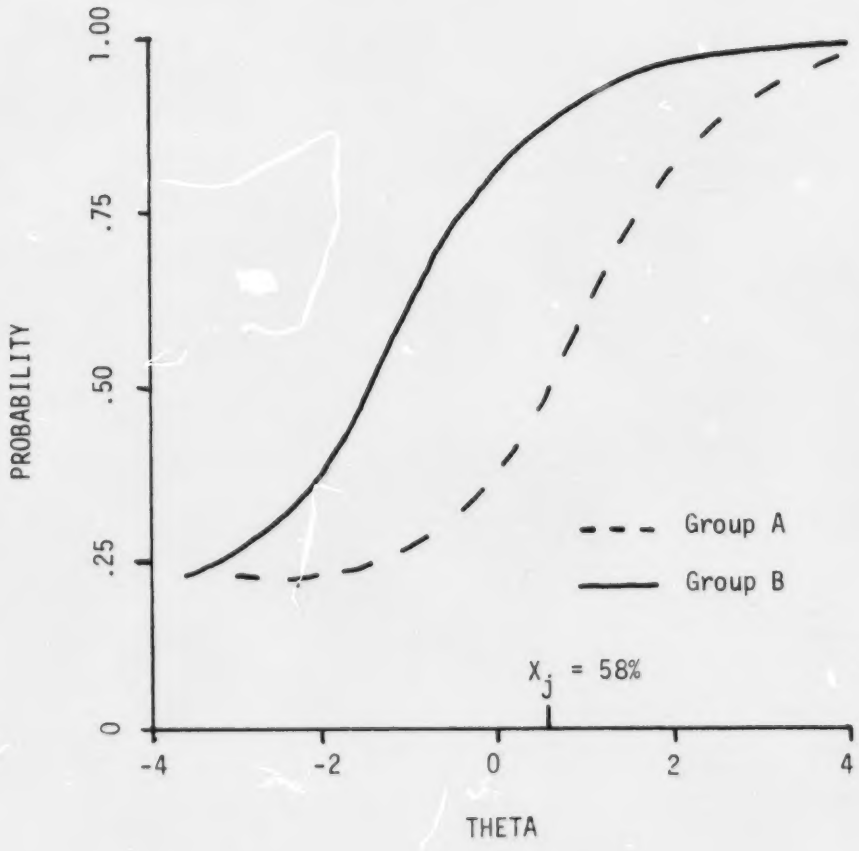


Figure 3. Two hypothetical response distributions.

Methodology

Two hypothetical non-biased tests were generated using the LTM approach to determine whether the TID approach might be overly sensitive in identifying biased items. Test 1 contained normally distributed item parameters while Test 2 contained unformily distributed item parameters. Parameters were choosen for both to simulate responses on a typical standardized test.

Two additional tests were generated to simulate a testing situation where items were obviously biased against a specific group of examinees. Item parameters were normally distributed for Test 3 and uniformly distributed for Test 4. Since the purpose of our research was to see how effectively the TID approach identified known biased items, a statistical test was used to measure the degree of congruence. A Pearson Product Moment correlation was computed using the bias values as defined by the LTM approach and the residual distances from the TID approach. Generally, the higher the correlation, the more the congruence between the two approaches.

Results

Item parameter estimates and the total test statistics are given in Table 1. The similarity of the means and standard deviations for both groups on each of the non-biased tests indicates the fairness of the tests for both groups of examinees.

In examining the total test statistics given in Table 1 for each of the biased tests, it is obvious that both the tests were biased against Group 2. Using the criterion of an area difference between the icc's for each group of equal to or greater than .70, the test with a normal distribution of item

Table 1

ESTIMATED PARAMETERS AND TOTAL TEST STATISTICS

	Test 1 No Bias Normal Distribution Mean		Test 2 No Bias Uniform Distribution Mean		Test 3 Some Bias Normal Distribution Mean		Test 4 Some Bias Uriform Distribution Mean	
Estimated Parameters:								
Difficulty	.03		.10		.03		.10	
Discrimination	1.00		1.00		1.00		1.00	
Guessing	.20		.20		.20		.20	
Diff. in Difficulty (Bias)	0.00		0.00		.33		.59	
Total Test Statisitcs:	GP1	GP2	GP1	GP2	GP1	GP2	GP1	GP2
Mean	36.00	35.79	34.96	34.78	38.06	33.82	38.30	31.53
Standard Deviation	11.04	11.34	11.93	11.81	10.77	11.26	11.93	11.43
Kuder-Richardson (KR-20)	.92	.92	.93	.92	.92	.92	.93	.92

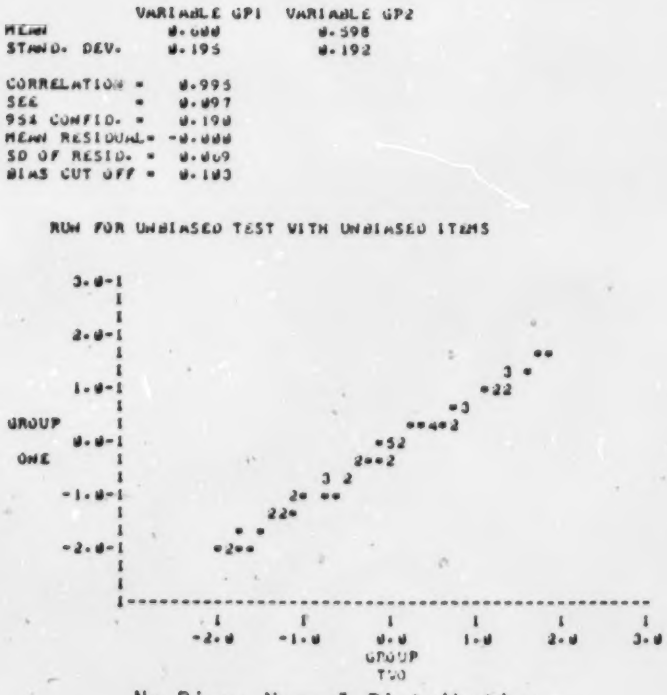
parameters identified seven items as being biased while the test with a uniform distribution of item parameters identified fourteen items as being biased. Items identified as biased in each of the generated tests were further described as the known biased items.

For the TID approach, using each of the criterion described earlier in this paper, the biased items were identified. Figure 4, shows the scatterplot of the transformed item p-values for each test. Theoritically, no biased items should have been identified in either of the non-biased tests. However, using the Strassberg-Rosenberg and Donlon criterion six items were identified as biased on Test 1, and eight items were identified as biased on Test 2. When the 95% confidence limit criterion was applied, no items were identified as biased for either test.

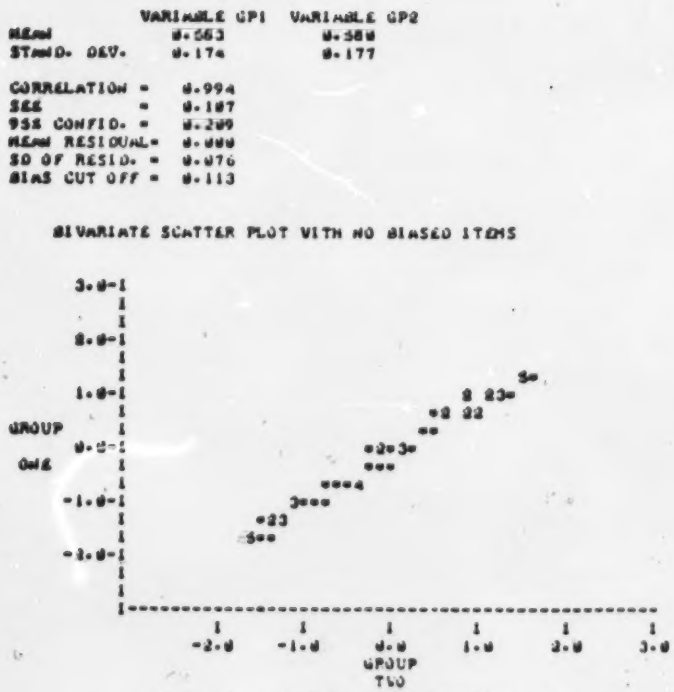
As in the non-biased test, biased items were identified under the TID approach using the two criterion discussed earlier. Figure 5 shows the scatterplot of the transformed item p-values for each test. For Test 3, six items were identified as biased using the Strassberg-Rosenberg and Donlon criterion and only one item using the 95% confidence limit criterion. For Test 4, the Strassberg-Rosenberg and Donlon criterion identified eight items as being biased. The 95% confidence limit criterion identified no biased items. However, our research was not primarily concerned with how many items were defined as biased under the TID approach. Rather, our research was more concerned with how effectively the TID approach identified the known biased items. Correlations were computed on each test to determine the congruence of the two approaches. The resulting correlations were:

$r = .95$ for Test 3; and

$r = .90$ for Test 4.



No Bias: Normal Distribution



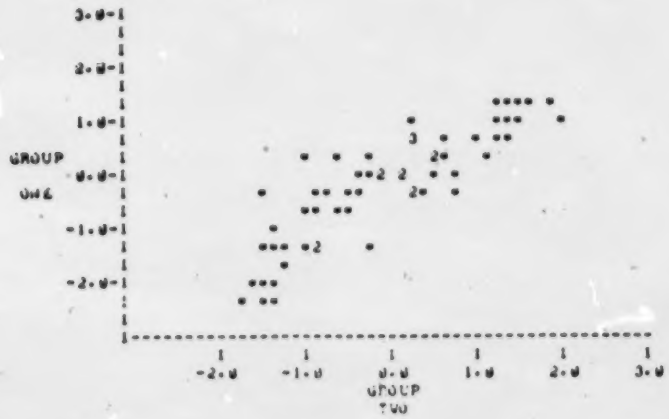
No Bias: Uniform Distribution

Figure 4. Non-Biased Tests Scatterplots

	VARIABLE GP1	VARIABLE GP2
MEAN	0.635	0.504
STAND. DEV.	0.202	0.193

CORRELATION =	0.874
SEE =	0.486
95% CONFID. =	0.952
MEAN RESIDUAL =	0.000
SD OF RESID. =	0.355
BIAS CUT OFF =	0.532

RUN FOR A BIASED TEST WITH SEVERAL BIASED ITEMS

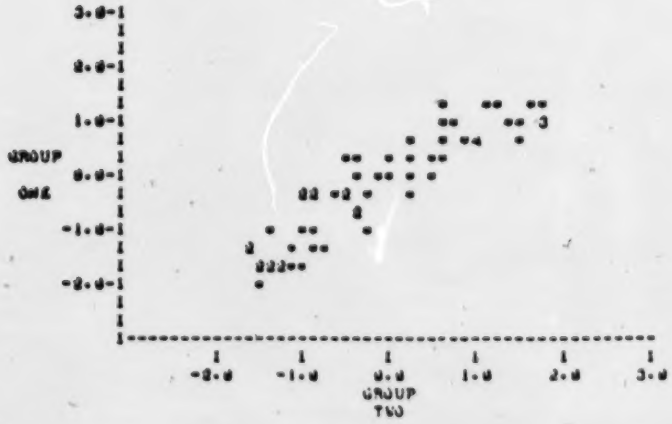


Bias: Normal Distribution

	VARIABLE GP1	VARIABLE GP2
MEAN	0.639	0.526
STAND. DEV.	0.178	0.179

CORRELATION =	0.924
SEE =	0.382
95% CONFID. =	0.749
MEAN RESIDUAL =	-0.000
SD OF RESID. =	0.275
BIAS CUT OFF =	0.413

BIVARIATE SCATTER PLOT FOR SEVERAL BIASED ITEMS



Bias: Normal Distribution

Figure 5. Biased Tests Scatterplots

From these high correlations, one would expect the TID to identify many of the same items as being biased. However, from examining Table 2, this is not the case. For Test 3, four of the seven biased items were found, and for Test 4, only six of the fourteen biased items were found. This discrepency seems to be the result of an inadequate criterion to detect biased items when the TID approach is applied.

Discussion

For the two biased tests, discrepencies existed in both the number of items identified as biased as well as which items were identified as being biased. In Test 3, using a criterion of .70 as a significantly biased item, the LTM approach identified seven items as biased. The TID approach correctly identified four items, failed to identify three items and falsely identified two items as being biased. In Test 4, the LTM approach identified fourteen items as biased. The TID approach correctly identified six items, failed to identify eight items and falsely identified two items as biased.

The strength of congruence was determined using the Pearson Product Moment correlation. For Test 3 and Test 4, the results were, respectively, $r = .95$ and $r = .90$. These results indicate a slightly lesser degree of congruence in the uniformly distributed test. However, it is the opinion of the authors that the lower correlation was more a function of the greater number of biased items (twice as many) rather than the type of item parameter distribution used. Research is in progress to confirm this opinion.

The global purpose of this study was to see if the TID approach would adequately identify items as biased that were known to be biased under the

Table 2

BIASED TESTS: IDENTIFICATION OF BIASED ITEMS[1]

Item #	Test 3	Test 4
2		+
4		*
5		-
8	*	
13		*
19		+
21		*
25	+	
27	-	*
28		+
30		*
33		+
38	+	
39		*
40		+
41	+	
44		*
45		-
48	-	
50	*	
51	+	+
58	*	*

+=TRUE

*=FAILURE

-=FALSE

[1]Only the items identified as biased by either the LTM or TID approach are listed.

LTM approach when ideally controlled conditions were used. For this study the ideal conditions consisted of: a 60 item test; a large number of examinees in each group (1000); discrimination and guessing parameters held equally distributed for each group within a narrow range of distribution; normal and equal distributions of ability for each group; and a known amount of bias introduced into the test only through the difficulty parameter. It was decided to look at both normally and uniformly distributed item parameters because of a study conducted by Pine and Weiss (1976). They used only the LTM approach and concluded that a uniform distribution of item parameters yielded a more valid test. However, the validity of our tests was not studied.

In these ideally controlled conditions, the TID approach did not adequately identify the known biased items when either of the two criteria were used. However, as discussed before, correlation analysis identified a high degree of congruence between the two approaches. It seems, then, that the inadequacy of the TID approach is a result of the lack of a good criterion for identifying biased items. When the Strassberg-Rosenberg and Donlon criterion was applied to either of the non-biased tests, several items were identified as biased; and too few items were identified as biased when it was applied to the biased test. The 95% confidence limit criterion worked adequately on the non-biased tests; but was totally inadequate when applied to the biased tests. This evidence tends to indicate that neither of the two criteria are functioning consistently. Consequently, neither can serve as a reliable criterion when applied to a standardized test where bias has not been previously controlled for. Therefore, consistent criteria needs to be developed for the TID approach.

Since standardized tests are being used more and more frequently in all

aspects of the educational environment and many diverse cultured groups are claiming that those tests are biased against them, it has become necessary to develop methods of screening the tests for bias before they are administered on a mass scale. However, research into ways of detecting item bias in these tests has been scattered and non-specific and often times limited to the study of two or three approaches. Many times the researcher has focused his research on the wrong issues. Researchers have developed their own methodology without adequately comparing it to methods used by others. If and when the method is tested against another method, the data sets frequently have not been similar enough to draw decisive conclusions.

Because of the above mentioned problems, the authors of this study are currently participating in another study to test the accuracy and robustness of six of the most commonly used item bias identification approaches using one data set to analyze each approach. Using a Monte Carlo approach, the conditions of each test can be fully controlled and an analysis made as to how each method functions. The authors feel that in conducting this research, a contribution can be made to aide the educational practitioner in gaining knowledge of which approach(s) are most suitable, most effective, most accurate and most reliable under given conditions.

REFERENCES

Angoff, W.H. A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu, May 1972.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.

Groome, M.L. A comparison of two approaches for item bias identification. Unpublished masters thesis, California State University, Sacramento, 1977.

Groome, W.R. Item bias: A comparison of two approaches. Unpublished masters thesis, California State University, Sacramento, 1977.

Lord, F.M. A theory of test scores. Psychometric Monograph Number 7. Princeton: Educational Testing Service, 1952.

Lord, F.M. and Novick, M.R. Statistical Theories of Mental Test Scores (2nd Ed.). Reading, MA: Addison-Wesley, 1974.

Pine, S.M. and Weiss, D.J. Effects of item characterisitcs on test fairness. (REsearch Rep. 76-5). Minneapolis: University of Minnisota, Department of Psychology, Psychometric Methods Program, December 1976. (AD A035393)

Rudner, L.M. An evaluation of select approaches for biased item identification. Unpublished doctoral dissertation, Catholic University, 1977.

Strassberg-Rosenberg, B. and Donlon, T.F. Context influences on sex differences in performance and aptitude tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., 1975.